蛋白质结构预测：梦想与现实

韦祎 杨继爽 袁雄鹰 邵明富 王超 卜东波

摘要：生物信息学的贡献可以从两个层面来衡量，一方面是对生物学的贡献，即能否以数理的手段辅助生物学家（或者独立地）做出新的生物学发现 (discovery)；另一方面则是对计算机科学的贡献，即实际问题是算法研究的推动力和源泉，我们能否在解决实际问题的过程中，检验已有算法，并开发新的算法(algorithm)。本文的目的即是以蛋白质结构预测的 FALCON 方法为样本，对上述两个层面的贡献做一阐述。简短地说，从生物学发现角度来讲，FALCON 的结果为“蛋白质结构构象数目是有限的”这一论断提供了定量的支持；从算法的角度讲，FALCON 实际上是一种新的优化框架，FALCON 能够大大降低搜索空间的大小，而经典的 Monte Carlo 与 Local search 始终维持一个相对较大的搜索空间。实验结果表明这种降低搜索空间大小的技术能够有效地提高搜索成功的可能性。

关键词：蛋白质结构预测；优化；原始对偶；采样

蛋白质是由肽键连接的一条氨基酸长链，只有在折叠成特定的形状之后方能产生特定的生物学功能。比如，疯牛病的病因即是脑部一种蛋白质朊蛋白(Prion Protein, PrP)结构发生变异：由正常的水溶性 α 螺旋结构，变异为不溶于水的 β 片状结构，从而沉积在脑组织中，引起神经细胞退行性改变，造成海绵状脑病。因此，了解蛋白质的结构对于认识蛋白质的功能有着重要意义。

使用生物学手段测定蛋白质三级结构的方法主要包括 X-晶体衍射实验和核磁共振（NMR）等，但以上两种蛋白质结构测定方法的速度远远跟不上 DNA 测序以及基因预测的速度，因而无法满足蛋白质组规模上（proteome-scale）结构预测的需求。比如，使用核磁共振方法测定一个蛋白质通常需要 15 万美元以及半年的时间。因此，人们希望使用计算进行预测来填补结构测定速度与序列测定速度之间的鸿沟。此外，预测方法的进展也有助于对蛋白质折叠机理的认识，从而具有重要的理论价值。更进一步，结构预测对于新蛋白质设计有着根本性的意义---要设计出具有某种特定结构的新蛋白质，结构预测无疑是缩短设计过程的一件利器。从这三个角度来说，从序列出发准确地预测蛋白质结构已成为人们的迫切要求。

那么，使用计算的方法从序列预测结构是可行的吗？

1965 年，安芬森（Anfinsen）基于还原变性的牛胰 RNase[1]在不需其他任何物质帮助下，仅通过去除变性剂和还原剂就使其恢复天然结构的实验结果，提出了“多肽链的氨基酸序列包含了形成其热力学上稳定的天然构象所必需的全部信息”的“自组装学说”，随后这个学说又得到一些补充。这些学说表明：氨基酸序列确定其空间构象，从而为蛋白质结构预测提供了可行性。

为客观、公正地衡量各种预测方法的性能，自 1994 年开始，莫尔特（John Moult）等人组织了一系列蛋白质结构预测技术评估（Critical Assessment of Techniques for Protein Structure Prediction，CASP）竞赛。和 Livebench 等其他测试方法不同，CASP 比赛采用了盲试（Blind Test）方法，即在每次比赛中使用的目标蛋白质的结构是未测定的，或者即使测

[1] 一种核糖核酸内切酶

定但是还未公开发布的。CASP 比赛在 2006 年度 CASP-7 比赛中共使用了超过 100 个测试用例，为算法设计和检验提供了一个比较公平的评测标准数据集。

值得指出的是，CASP 比赛的目的是促进新思路的产生，而不是简单地评价现有的各种方法、实现的好坏等等。这也许是我们看待各种国际比赛的最佳态度。

经典的蛋白质结构的预测方法可以分为三类，即：同源建模方法（Homology Modeling），穿线法(Threading)和从头预测（ab initio）方法。

同源建模方法的核心思想是通过目标序列的同源蛋白质来推定其三维结构。其关键步骤是序列-序列（sequence-sequence）相似性比较，以推断蛋白质之间的同源关系。对于相似度比较高的情况，同源建模方法能够以很高的精度预测出蛋白质三级结构；而在序列相似度较小的情况下则往往失效。

穿线法的核心思想是寻找和目标序列没有显著性同源关系、但是具有同一结构折叠（Fold）类型的蛋白质。其关键步骤是序列-结构比较计算，以获得最可能的比对（alignment）。和同源建模方法相比，穿线法的主要不同是充分利用了模板库中的结构信息，比如氨基酸之间的相互作用等。因此，穿线法能够得到比同源建模更精确的预测结果。

从头预测方法的核心思想是从第一性原理出发，寻找目标蛋白质能量最小的构象。IBM 的超级计算机 BlueGene-L 就是为了实现这个模拟过程而研制开发的，但是目前只能计算几个氨基酸的折叠过程；采用蒙特卡罗（Monte Carlo）策略，段（Duan）和科尔曼（Kollman）在 256 个处理器的克雷（Cray）机器上计算了两个月，仅仅模拟了 36 个氨基酸的一毫秒的真实折叠过程。

经过多年的努力，现在对于序列相似度大于 30%的同源蛋白质来说，结构预测问题可以认为已经解决；穿线法识别折叠类型的准确率大概为 2/3；而从头预测方法还需要大量的努力和新思路才能取得突破。

近年来，从头预测方法得到越来越多的重视，其原因在于和同源建模以及穿线法相比较而言，从头预测方法具有其独特的优势，比如有助于揭示蛋白质折叠机理，能够在同源蛋白质未知的情况下预测结构等。但是该方法也存在一些不足，比如研究人员通常使用简单枚举的“离散”方式来描述局部结构的多个候选构象，而不是刻画“连续”构象空间的分布，造成每个候选虽然和真实结构很相似，但是仍然存在较大的误差，而且这种误差无法消除。从头预测遇到这种局部结构的离散性往往无能为力；此外搜索空间过大也是一个问题，直接导致搜索到真实结构的概率大大降低。上述不足之处造成了从头预测方法在实际应用中的困难。针对这些不足，我们认为有必要设计新的算法框架。

作为前驱性工作，李帅诚、卜东波、许锦波、李明提出了一种基于 Fragment-HMM[2] 的新预测算法 FALCON[3]，能够将蛋白质结构构象空间大小从 ROSETTA 方法的 $\boldsymbol{O}(200^n)$ 降低至 $\boldsymbol{O}(1.66^n)$，从而更接近于迪尔（K. Dill）的估计值 $\boldsymbol{O}(1.6^n)$。

我们的方法的生物学依据是：蛋白质结构是由近程相互作用和远程相互作用共同作用的结果，蛋白质局部结构主要受近程相互作用影响，而远程相互作用则影响各个局部结构的摆放位置，使其自由能最小，从而产生稳定结构。因此，我们就要解决以下两个问题：

---

2 Fragmentation hidden Markov model 针对片段的隐马尔科夫模型

3 卜东波于加拿大滑铁卢大学访问期间与李帅诚、许锦波、李明教授共同完成。

1. 如何刻画局部的结构倾向性？

2. 如何刻画远程相互作用导致的相关性？

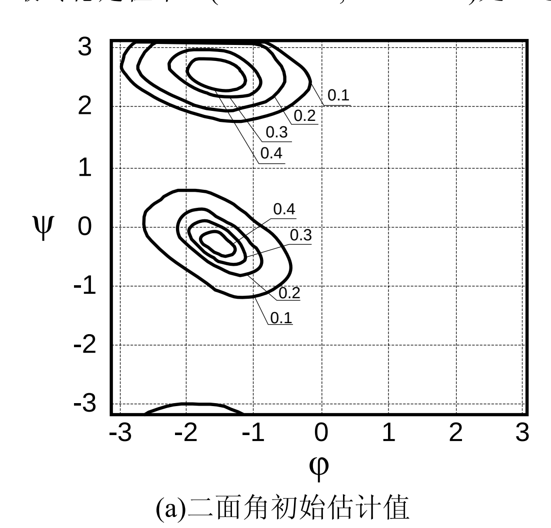
我们开发的 FALCON 算法采取了如下技术：

1. 局部结构（Local Structure）的预测算法：

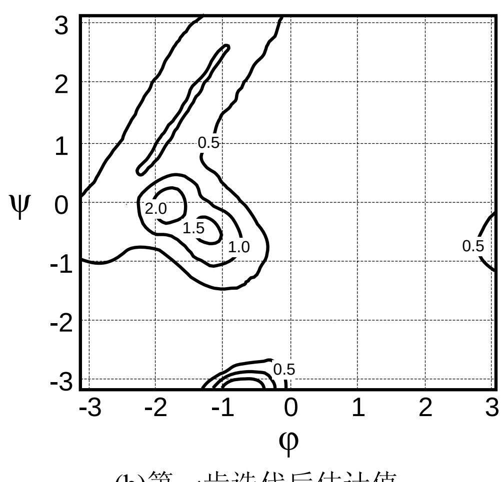
我们初步实现了上述算法，开发了软件包 FRazor (Fragment Razor)，并获得一些初步实验结果。在实验中，我们从 PDB[4]模板库中随机抽取了 9338 个片断，一半作为训练集，一半作为测试集合。初步实验结果表明：如果设置局部结构候选集合规模为 25，则我们的整数线性规划模型对于 Alpha 螺旋（Alpha helix）区域的命中率是 98.6%，β链（Beta Strand）区域是 89.6%，环（Loop）区域是 78.1%。这比 ROSETTA 方法的结果有较大的改进。如果设置候选集合的规模为 40，则命中率分别为 99%，92.9%，和 82.4%。这表明这种整数线性规划模型能够有效地预测出局部结构。

2. 二面角分布刻画与逐步求精：

我们初步实现了迭代策略，实验结果初步表明其有效性。下图显示的是对蛋白质 2CRO（Cro Repressor）的 Residue 41 的二面角估计值随着迭代进行不断改进的情况。对这个残基来说，通过局部结构预测步骤产生的初始估计值可以分作两个团，一个位于 Alpha 螺旋区域，一个位于β链区域，然而这两个团和真实值（φ=1.44, ψ=-0.63）都有较大差异；经过一步迭代之后，β链团和 Alpha 螺旋团都变弱，而新出现了一个估计值集中区域（中心点 $\varphi=-1.82$， $\psi=-0.07$）；再经过第二步迭代之后，错误的β链团彻底消失，而 Alpha 螺旋团继续变弱；经过第三步迭代之后，Alpha 螺旋团也最终消失了，新出现的团逐步变强，最终稳定在中心($\varphi=-1.86$, $\psi=-0.13$)处。这个中心和真实值比较接近，相应于环结构。

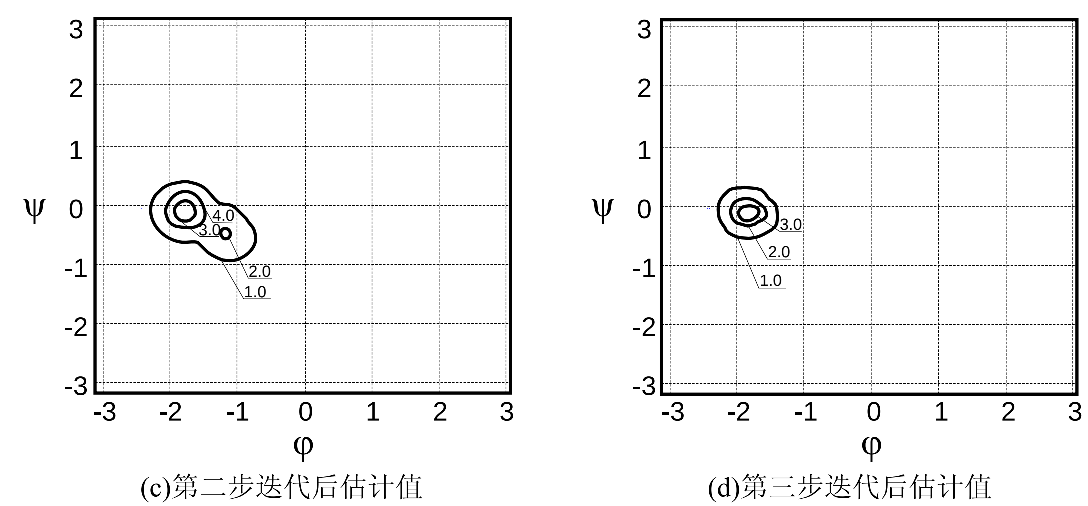


(a)二面角初始估计值



(b)第一步迭代后估计值

[4] Protein Data Bank，蛋白质数据库



(c)第二步迭代后估计值 (d)第三步迭代后估计值

图1. Residue 41 的二面角估计值随着迭代进行不断改进

3. 基于位置特异性隐马尔科夫模型的采样算法：

和 FB5-HMM 等已有算法不同的是，我们设计的 FALCON 是一种位置特异性隐马尔科夫模型（Position-specific HMM），即每个位置上的隐含结点数目和转移概率都不同。实验结果表明：这种位置特异性隐马尔科夫模型能够有效地降低搜索空间。

我们已经初步实现了上述模型，开发了软件包 FALCON 的初步原型。实验结果表明，即使不采用迭代技术，该算法就已经表现出对 ROSETTA 的优势。

目标蛋白质	ROSETTA		FALCON	
	Best	<6.0Å(%)	Best	<6.0Å(%)
Protein A, 1FC2	2.82	80.2	2.64	94.3
Homeodomain，1ENH	1.52	94.4	1.81	92.8
Protein G, 2GB1	2.21	53.7	2.18	93.4
Cro repressor, 2CRO	2.56	70.4	2.48	75.8
Protein L7/L12, 1CTF	1.44	14.3	0.56	25.6
Calbidin, 4lCB	3.87	19.9	2.93	46.3

如果采用迭代技术，FALCON 的结果会大大改善。经过 5 轮迭代之后，对于这 6 个测试蛋白质用例而言，其“好结构”的比例都可以逐步提高到 100%。

目标蛋白质	迭代次数					
	1	2	3	4	5	6
Protein A, 1FC2	94.3	98.5	100	100	100	100
Homeodomain，1ENH	92.8	95.0	96.9	100	100	100
Protein G, 2GB1	93.4	96.4	100	100	100	100
Cro repressor, 2CRO	75.8	97.3	100	100	100	100
Protein L7/L12, 1CTF	25.6	68.8	97.0	100	100	100
Calbidin, 4lCB	46.3	90.5	99.3	100	100	100

在 CASP-8 比赛中，FALCON 获得了折叠识别困难类（Fold Recognition Hard）的第三名。

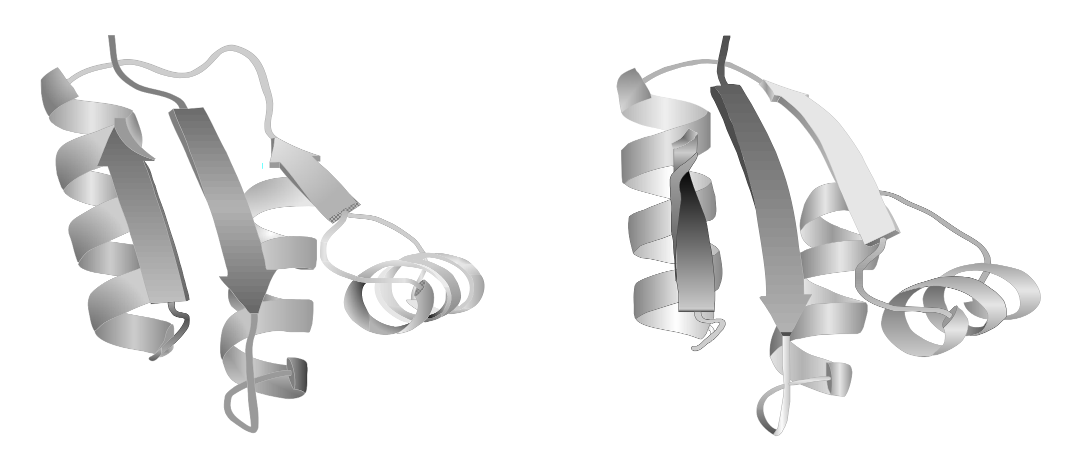


图2. 对蛋白质 1CTF 的预测结果（左）与其天然结构（右），误差 0. 557 埃（Å）。

本质上，FALCON 是将传统的离散域优化问题转化成如下的连续域优化问题：

$$\begin{aligned} &\boldsymbol{min}\, E(\varphi_1,\psi_1,\cdots,\varphi_n,\psi_n) \\ &s.t. \quad (\varphi_i,\psi_i) \sim f_i \end{aligned}$$

其中 $\phi_i \in [-\pi,\pi], \psi_i \in [-\pi,\pi]$ 是角度变量，以表示组成蛋白质的第 i 个氨基酸的两个二面角。确定出所有位置氨基酸的二面角，就能够精确地恢复出整体空间结构。f_i 表示 (ϕ_i,ψ_i) 在连续空间上的分布，目标函数 E 表示蛋白质在由当前二面角确定的结构构象下的能量。整个优化问题目标就是使用采样技术(sampling)最终求解上述优化问题。

值得指出的是：在经典的蒙特卡洛或局部搜索方法中搜索空间不变，而我们的算法能够大幅地缩小搜索空间。

我们的体会是：对于优化问题，问题变换是改进算法的重要手段，即要么改变搜索空间，要么改变能量图景（energy landscape）。

虽然 CASP-8 比赛结果表明 FALCON 作为一个原型系统取得了初步成功，但要真正达到“无缝衔接同源建模、穿线法和第一性技术”的理想目标，还有很多理论和实践上的困难需要克服。我们依然在努力。

作者简介：
韦 祎 中国科学院计算技术研究所 2006 级硕士研究生
杨继爽 中国科学院计算技术研究所 2006 级硕士研究生
袁雄鹰 中国科学院计算技术研究所 2007 级硕士研究生
邵明富 中国科学院计算技术研究所 2008 级硕士研究生
王 超 中国科学院计算技术研究所 2008 级硕博连读生
卜东波 中国科学院计算技术研究所 副研究员 dbu@ict.ac.cn